Reward Shaping for Robust Refusal in Small Language Models for Retrieval-Augmented Question Answering

Thilina C. Rajapakse*
Eindhoven University of Technology
Eindhoven, The Netherlands
t.c.r.rajapakse@tue.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Abstract

We focus on smaller open-source LMs (2–7B parameters), which are attractive for practical deployment due to their lower computational cost and greater accessibility than frontier-scale models. We show that instruction-tuned models generate answers even when explicitly prompted to refuse when the answer is not supported by the documents. In the presence of distractor documents, instruction-tuned models demonstrate inconsistent performance, with answer accuracy metrics deteriorating in most cases. To mitigate this behavior, we introduce Reward Shaping for Refusal and Reasoning (RSRR), a reinforcement learning framework that teaches LMs to reason step-by-step over multiple documents and to refuse to answer when evidence is insufficient. Models trained with RSRR achieve substantial improvements in robustness to distractor documents and in correct refusal accuracy, with gains of 39.8% and 43.3%, respectively. We release code and data to reproduce all results.[1]

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; **Language models**; **Question answering**.

Keywords

RAG, Question answering, Language models.

ACM Reference Format:
Thilina C. Rajapakse and Maarten de Rijke. 2026. Reward Shaping for Robust Refusal in Small Language Models for Retrieval-Augmented Question Answering. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3809891

1 Introduction

Retrieval-augmented generation (RAG) extends language models (LMs) knowledge beyond training [4, 6, 12], which is essential for continuously updated information where retraining is infeasible [4, 8]. In high-stakes domains, ensuring truthfulness is critical [3, 7, 26, 27]; therefore, the LM must generate responses based solely on the retrieved documents. We focus on the behavior of the LM component of a RAG system and analyze the ability of open-source LMs to (a) refuse to answer queries when sufficient evidence is not present in the provided information, (b) accurately answer complex queries given sufficient information (in one or more documents), and (c) robustly answer complex queries in noisy retrieval scenarios with relevant and non-relevant documents.

Small LMs. Small open-source LMs remain highly relevant [1, 2, 30]. They are efficient to train and deploy, making them practical in domains where computational budgets, privacy requirements, or regulatory constraints preclude the use of frontier models [23, 25, 28]. Small models are easier to steer with explicit reward shaping, providing a tractable testbed for studying how refusal behavior can be trained rather than emerged implicitly [14, 24].

Observed behavior. Our analysis shows that instruction-tuned LMs struggle with three key patterns: (a) they hallucinate answers instead of refusing when evidence is missing; (b) they achieve only moderate accuracy in vanilla settings; and (c) they are highly sensitive to distractor noise. Ultimately, prompting and instruction-tuning alone cannot ensure reliable refusal or robustness in noisy, real-world RAG scenarios.

Research goal. The primary research question we pursue in this paper is how relevance-based rewards can be used in reward shaping to train small language models to answer based on retrieved evidence and to refuse when evidence is insufficient.

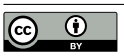
Our approach. We introduce *reward shaping for robust refusal* (RSRR), a reinforcement learning framework that trains models to both ground their answers in evidence and to refuse to answer when evidence is insufficient. We extend proximal policy optimization (PPO) [13, 18, 31] with reward components tailored for retrieval-augmented QA [4, 6, 12, 19]: (a) a *relevance-based reward* that encourages models to generate reasoning more relevant than any single input document, (b) an *answer correctness reward* that schedules a penalty/bonus depending on whether the gold answer (including `NO-RES`) is present, (c) a *formatting reward* that enforces structured outputs, and (d) a KL penalty to stabilize training.

Contributions. (a) *Evaluation*: We show that small instruction–tuned LMs struggle with calibrated refusal and robustness across vanilla, refusal, and distractor settings. (b) *RSRR framework*: We introduce a reward-shaping framework that couples grounded generation with explicit refusal training. (c) *Results*: RSRR substantially improves refusal accuracy (+43.3% relative) and distractor robustness (+39.8% relative), making small LMs more reliable for RAG.

2 Related Work

Sensitive domains. Hallucination mitigation and calibrated refusal is especially important in high-stakes domains such as healthcare. Recent work [11, 20] underscores the prevalence of hallucination,

*Work conducted at the University of Amsterdam.
[1] https://github.com/ThilinaRajapakse/rsrr



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809891

Table 1: Dataset statistics for evaluation.

Dataset	Test size	Answer type
BioASQ	2,688	yes/no, span
PubMedQA	500	yes/no/maybe
BeerQA	14,121	span
StrategyQA	229	yes/no

especially in the healthcare domain, and focuses on hallucination detection or reduction via better retrieval. In contrast, we examine the behavior of LMs *after* the retrieval step. We evaluate the QA accuracy of LMs under ideal and noisy conditions, as well as the ability to *refuse* when sufficient evidence is absent, complementing retrieval-centric approaches with mechanisms for conservative abstention.

Refusal. Recent work uses RL and reward models to align LM refusal [17, 27], though naive optimization can cause over-refusal [22]. Unlike studies focusing on internal knowledge, we emphasize evidence-based refusal. By shaping rewards to favor abstention for unanswerable queries and faithful reasoning for answerable ones, we ensure the calibrated refusal necessary for reliable RAG.

Reward shaping and alignment. While prior work shapes reward functions for broad alignment, such as factuality, style, or safety [17, 29], we focus on coupling evidence-based reasoning with *refusal*. By rewarding multi-document synthesis alongside conservative abstention, we address a key RAG gap: ensuring models generate answers based strictly on provided evidence and recognize when that evidence is insufficient.

3 Analysis of Instruction-Tuned Models

We evaluate instruction-tuned models on their ability to answer based on evidence or refuse when it is missing. To isolate the QA component from retriever noise, we provide "gold" documents in two of our three settings. We analyze: (a) *vanilla*, testing reasoning on complex queries; (b) *refusal*, testing whether models decline to answer when evidence is absent; and (c) *distractors*, where we include pseudo-relevant documents to simulate an imperfect RAG retriever.

Models. We test four instruction-tuned, open-source language models: Llama 3, Gemma 2, Qwen 2.5, and DeepSeek-R1. We use instruction-tuned versions of the models with no additional training.

Datasets. We evaluate instruction-tuned models on four datasets (Table 1), using provided evidence to simulate RAG scenarios. *PubMedQA* [9] is a biomedical dataset requiring reasoning over context for yes/no/maybe answers. *BioASQ* [10] provides biomedical yes/no and factoid questions for exact-match evaluation. *BeerQA* [21] and *StrategyQA* [5] both require multi-hop reasoning; BeerQA uses Wikipedia passages, while StrategyQA focuses on inferring implicit reasoning steps.

Augmentation. Beyond the *vanilla* setting (all queries answerable), we introduce two augmented settings (Section 5): (a) *Refusal (No-Res)*: Original evidence is replaced with pseudo-relevant documents to test if models decline to answer when evidence is insufficient. (b) *With Distractors*: To simulate imperfect retrieval, we provide a mix of relevant and distractor documents, testing model robustness to noise.

Table 2: Tagged accuracy of instruction-tuned models in vanilla, distractor, and No-Res scenarios (mean across prompt lengths). (PMQA is short for PubMedQA, StratQA for StrategyQA, and Distr. for Distractors.)

Setting	Model	BioASQ	BeerQA	PMQA	StratQA	Avg.	Wins
No-Res	R1-7B	0.308	0.663	0.408	*0.699*	0.520	0
	Gemma-2-2B	0.479	**0.800**	0.562	0.472	0.578	1
	LLaMA-3.2-3B	*0.484*	*0.682*	*0.723*	0.617	0.627	0
	Qwen-2.5-3B	**0.683**	0.662	**0.921**	**0.927**	0.749	3
Vanilla	R1-7B	0.402	*0.588*	**0.485**	**0.460**	0.484	2
	Gemma-2-2B	**0.477**	0.474	0.340	0.138	0.357	1
	LLaMA-3.2-3B	0.321	**0.596**	*0.464*	0.351	0.433	1
	Qwen-2.5-3B	*0.402*	0.453	0.392	*0.358*	0.401	0
Distr.	R1-7B	0.329	*0.497*	**0.390**	**0.458**	0.419	2
	Gemma-2-2B	**0.482**	0.491	0.233	0.287	0.373	1
	LLaMA-3.2-3B	0.323	**0.542**	*0.366*	0.368	0.400	1
	Qwen-2.5-3B	*0.383*	0.355	0.235	*0.373*	0.336	0

Evaluation. We evaluate models across three scenarios: No-Res (evidence removed; target is NO-RES), Vanilla (gold evidence provided), and Distractor (gold evidence plus irrelevant documents). We report *tagged accuracy*, requiring answers within `<ANSWER>...</ANSWER>` tags to match a gold reference after standard normalization (lowercasing, punctuation removal, and trimming). This metric assesses grounding and the ability to refuse unanswerable queries. Scores are averaged across three prompts (see repository, Footnote 1).

Results of the analysis. Table 2 reveals three consistent weaknesses in small instruction-tuned LMs: unreliable refusal, limited reasoning, and sensitivity to noise. While Qwen-2.5-3B exhibits the strongest refusal behavior, most models hallucinate when evidence is absent, proving that prompting alone is insufficient for calibration. In the *vanilla* setting, performance is only moderate, as these models struggle to integrate information across documents. The addition of *distractors* further degrades performance (except for Gemma-2-2B). These limitations highlight that small LMs cannot yet reliably handle RAG tasks through prompting alone, motivating our reward-shaping approach to improve faithfulness and refusal.

4 Reward Shaping for Refusal and Reasoning

Proximal policy optimization. We adopt proximal policy optimization (PPO) as the reinforcement learning algorithm for fine-tuning instruction-tuned language models with our proposed reward signals. PPO provides a stable and sample-efficient method for policy optimization while preventing large, destabilizing updates. Formally, given a policy π_θ parameterized by θ, PPO maximizes the clipped surrogate objective:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\Big], \quad (1)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio between the updated and old policies, and $\hat{A}_t$ is the estimated advantage at timestep

t. The clipping parameter ϵ prevents excessively large policy updates. In our setup: (a) the *policy* (π_θ) is the instruction-tuned LM generating step-by-step reasoning and final answers; (b) the *environment* is a query together with retrieved documents; (c) *actions* are token-level generations, terminated by end-of-sequence; and (d) *rewards* are shaped signals described below. Following standard PPO-RLHF practice, we add a scalar value head to the model to estimate state values for advantage computation. A KL penalty is applied to ensure the updated policy remains close to the reference model, preventing catastrophic drift during training.

Relevance-based reward. The central component of our reward design is a relevance-based signal that encourages model reasoning to be grounded in the retrieved evidence. For query q, model output y is parsed to obtain a reasoning span $r = \text{extract}(y, \texttt{REASONING})$. Let $\mathcal{D} = \{d_i\}_{i=1}^{K}$ be the set of documents provided to the model. We use a learned ranking model $Rel(q, \cdot)$ to score the relevance of the reasoning span and the documents to the query. We define the relevance reward as

$$R_{\text{rel}}(q, y, \mathcal{D}) \;=\; Rel(q, r) \;-\; \max_{d \in \mathcal{D}} Rel(q, d), \tag{2}$$

which is positive only if the model's reasoning surpasses the relevance of any single document. This promotes faithful multi-document integration and distractor resistance. We scale it by λ_{rel} in the final objective.

Alignment reward. We add an alignment reward R_{align} from a separate reward model to ensure fluency and coherence. This complements the relevance signal by ensuring outputs are both grounded and well-formed. Formally,

$$R_{\text{align}}(q, y) \;=\; \lambda_{\text{align}}\, g(q, y), \tag{3}$$

where g is the alignment reward model and λ_{align} is a weighting coefficient.

Answer correctness. To encourage inclusion of the correct answer, we add a scheduled shaping term. For each example, let $a^\star$ denote the gold answer or `NO-RES` if the query is unanswerable, and define $\text{correct}(y, a^\star) = 1$ if the extracted `ANSWER` span from y matches $a^\star$ under task normalization, else 0. The reward is

$$R_{\text{correct}}(y, a^\star, t) \;=\; \begin{cases} +\phi(t), & \text{correct}(y, a^\star) = 1, \\ -\phi(t), & \text{correct}(y, a^\star) = 0, \end{cases} \tag{4}$$

where $\phi(t) \geq 0$ grows over steps t: small early (encouraging exploration), larger later (enforcing correctness).

Formatting reward. Outputs must follow the required template with `<REASONING>` and `<ANSWER>` tags. Let $\text{valid}(y) \in \{0, 1\}$ indicate conformity. With a fixed $\gamma_{\text{fmt}} > 0$, we assign

$$R_{\text{fmt}}(y) = \begin{cases} +\gamma_{\text{fmt}}, & \text{valid}(y) = 1, \\ -\gamma_{\text{fmt}}, & \text{valid}(y) = 0. \end{cases} \tag{5}$$

KL. We regularize π_θ against a reference π_{ref} with a per-token KL term:

$$r_t^{\text{KL}} = -\beta\, \text{KL}\big(\pi_\theta(\cdot|s_t) \,\|\, \pi_{\text{ref}}(\cdot|s_t)\big), \tag{6}$$

where β controls penalty strength and prevents drift from the initialization.

Final reward. The overall reward combines all sequence-level terms with KL regularization. At the sequence level we define

$$R_{\text{seq}} = \lambda_{\text{rel}} R_{\text{rel}} + \lambda_{\text{align}} R_{\text{align}} + R_{\text{correct}} + R_{\text{fmt}}, \tag{7}$$

where λ_{rel} and λ_{align} weight the relevance and alignment components, and R_{correct}, R_{fmt} are scaled by their hyperparameters $\phi(t)$ and γ_{fmt}. Following standard RLHF practice, R_{seq} is added at the final token while per-token KL penalties r_t^{KL} are applied throughout:

$$\mathbf{r} = (r_1^{\text{KL}}, \ \ldots, \ r_{T-1}^{\text{KL}}, \ r_T^{\text{KL}} + R_{\text{seq}}). \tag{8}$$

This vector is used for advantage estimation with GAE and PPO optimization.

5 Experimental Setup for Reward Shaping

Datasets. We use the BeerQA dataset for training as it consists of multi-hop questions that require evidence from multiple documents to answer correctly. We augment the training set of BeerQA to add examples containing distractor documents along with the relevant documents, and we also generate unanswerable versions of each query by replacing all relevant documents with distractor documents; see below.

We evaluate on the QA datasets detailed in Section 3: PubMedQA, BioASQ, BeerQA, and StrategyQA. We cast each example as retrieval-augmented QA with evidence documents $\mathcal{D}$. We follow the evaluation procedure from Section 3 and evaluate the models under *refusal*, *vanilla*, and *with distractors* settings.

We use the official test or held-out evaluation splits when available and reserve a subset of the remaining data for training (Table 1).

Data augmentation. We augment each dataset to create three settings: (a) *NO-RES:* gold evidence is withheld; the correct target is `NO-RES`. (b) *Vanilla:* all gold evidence is present; queries are answerable. (c) *Distractors:* gold evidence is present *plus* k non-relevant documents sampled from a retrieval ranking (details below).

We sample k distractors per query from a ranked candidate list (excluding gold), using top-M as the pool, such that each example has n associated documents. The ranked candidate lists are built using BM25. To fit within the model's context length, we set $n \in [3, 20]$ and $M \in [10, 40]$, with $k = n -$ (number of relevant documents). Both n and M are held constant per dataset.

In the *refusal* setting, all gold documents are replaced with sampled distractor documents.

Prompt and output template. All models are prompted with a fixed instruction and the evidence block (`Documents:`), followed by `Question:` and a required output template enforced via the formatting reward (Section 4):

```
<REASONING> ... </REASONING>
<ANSWER> ... </ANSWER>
```

Models. We fine-tune an instruction-tuned backbone as the policy π_θ and use the same initialization as the reference π_{ref} for KL control (Section 4). We report results for *LLaMA-3.2-3B*. $g(q, y)$ is a sequence-level model scoring general quality (fluency, coherence, helpfulness, etc.). We use the publicly available model checkpoint

Table 3: Tagged accuracy of PPO-trained models under different QA scenarios (mean across prompt lengths). (Same conventions as in Table 2.)

Setting	Model	BioASQ	BeerQA	PMQA	StratQA	Avg.	Wins
No-Res	LLaMA-3.2-3B	0.483	0.683	*0.723*	0.617	0.627	0
	PPO - No Rank	**0.834**	**0.981**	**0.975**	**0.994**	0.946	4
	RSRR	*0.699*	*0.958*	*0.962*	*0.993*	0.903	0
Vanilla	LLaMA-3.2-3B	0.321	0.596	*0.464*	**0.351**	0.431	1
	PPO - No Rank	*0.497*	*0.634*	**0.477**	0.332	0.485	1
	RSRR	**0.610**	**0.753**	0.463	**0.351**	0.544	3
Distr.	LLaMA-3.2-3B	0.323	*0.542*	*0.366*	**0.368**	0.400	1
	PPO - No Rank	*0.483*	0.498	0.489	0.295	0.441	0
	RSRR	**0.611**	**0.678**	**0.507**	*0.345*	0.535	3

Skywork/Skywork-Reward-Llama-3.1-8B-v0.2 [15]. $Rel(q, \cdot)$ is a ranking model used to score (q, r) and (q, d_i) pairs for the relevance reward (Section 4). We use the publicly available model checkpoint *rankllama-v1-7b* [16]. We fine-tune LLaMA-3.2-3B on augmented BeerQA using PPO. We compare a PPO - No Rank baseline (standard PPO without relevance rewards) against RSRR, which incorporates our proposed relevance-based reward (Section 4). Both models use the hyperparameters detailed below.

Hyperparameters. All models were trained for one epoch on 8×NVIDIA L40 GPUs (48GB), corresponding to ∼2,090 PPO updates. We adopt the Hugging Face `trl` defaults for PPO unless otherwise noted. The policy π_θ was optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, $\epsilon = 10^{-8}$) at a learning rate of 3×10^{-6}. We run two PPO epochs with a single mini-batch split, for an effective batch size of 128, and set a maximum generation length of 500 tokens. For sequence-level shaping, we use $\lambda_{\text{rel}} = 1.0$ for R_{rel} and $\lambda_{\text{align}} = 0.8$ for R_{align}. The formatting reward R_{fmt} uses $\gamma_{\text{fmt}} = 1.0$, giving a symmetric bonus/penalty for valid vs. invalid outputs. The correctness reward R_{correct} is scheduled linearly from 3.0 to 7.0 between steps 100 and 800 ($\phi(t)$), encouraging exploration early and enforcing refusal later. A penalty of -2.0 is applied if the output does not end with an end-of-sequence marker. We retain the `trl` default clipping range of 0.2 for both policy and value updates.

6 Results for Reward Shaping

We compare the instruction-tuned baseline Llama model, PPO without the relevance-based reward (PPO - No Rank), and the full RSRR method across three settings: *Vanilla*, *Distractors*, and *No-Res*. Performance is measured using tagged accuracy (Table 3) and percentage of incorrect refusals (Table 4), capturing how often the model abstains despite sufficient evidence.

Table 3 reports tagged accuracy scores. In the *No-Res* setting, PPO - No Rank achieves the highest scores, but this is largely due to an over-refusal strategy: the model learns to abstain excessively, inflating refusal accuracy but undermining performance on answerable questions. RSRR performs slightly lower in raw accuracy but achieves more calibrated refusal. In the *Vanilla* setting, both PPO variants perform comparably to the baseline. RSRR improves performance on BioASQ and BeerQA, but shows drops on StrategyQA. Since StrategyQA often requires implicit multi-step reasoning whose evidence is not always directly expressed in the retrieved documents, RSRR's conservative bias can suppress otherwise valid answers. In the *Distractor* setting, RSRR consistently outperforms PPO - No Rank and the baseline. By rewarding reasoning that is more relevant to the query than any single distractor document, RSRR achieves better robustness to noisy retrieval, maintaining higher accuracy even in the presence of irrelevant evidence.

Table 4: Percentage of incorrect refusals. Lower values are better. Reported across datasets for the Vanilla and Distractor settings. (Same conventions as in Table 2.)

Setting	Model	BioASQ	BeerQA	PMQA	StratQA	Avg.
Vanilla	LLaMA-3.2-3B	**12.31**	**7.48**	**10.80**	**9.17**	9.94
	PPO - No Rank	27.38	11.35	40.60	64.19	35.88
	RSRR	*17.96*	*8.76*	*31.60*	*54.58*	28.22
Distr.	LLaMA-3.2-3B	**11.60**	**9.87**	29.20	**17.24**	16.98
	PPO - No Rank	30.09	28.70	36.80	92.13	46.93
	RSRR	*15.47*	*16.24*	**27.80**	*85.58*	36.27

To better diagnose errors beyond tagged accuracy, we report the percentage of *incorrect refusals*, i.e., cases where the model abstains despite sufficient evidence being present (Table 4). PPO - No Rank suffers from severe over-refusal, particularly on PubMedQA and StrategyQA, where it abstains in 40–90% of answerable cases. Reward Shaping for Refusal and Reasoning substantially reduces this failure mode, though it still exhibits higher incorrect refusal rates than the baseline. The baseline, in contrast, minimizes refusal errors but, as Table 3 shows, this comes at the cost of weaker robustness in distractor settings.

Overall, Reward Shaping for Refusal and Reasoning improves robustness in distractor settings and produces more calibrated refusals than PPO - No Rank, with some cost to raw accuracy. Our results show that relevance-based reward shaping, combined with scheduled correctness and format rewards, enables small LMs to reason faithfully and abstain when evidence is insufficient. This approach significantly improves refusal accuracy and distractor robustness, prioritizing reliability over mere accuracy. While RSRR can lower raw accuracy, particularly on StrategyQA, where implicit reasoning steps may trigger a conservative bias, we view this trade-off as intentional. In high-stakes domains, avoiding unsupported answers is often preferable to maximizing coverage, though balancing these remains a future challenge.

7 Conclusion

We introduced RSRR, a reward-shaping framework that trains models to ground answers in evidence and refuse when it is missing. It improves refusal accuracy and distractor robustness by combining a relevance-margin reward with scheduled correctness and format shaping. Our results show that small LMs can move beyond surface-level accuracy toward calibrated, trustworthy RAG behavior.

Reliable refusal is critical in high-stakes domains like healthcare and law to prevent harmful, overconfident errors. While our approach fosters trust by reducing misinformation, it risks "over-refusal" if penalties are poorly calibrated. This is a step toward trustworthy RAG, though practical deployment requires careful calibration to maintain the balance between caution and utility.

As to limitations, the reliance on *structured tags* and a fixed `NO-RES` string may not generalize to open-ended or conversational settings. Second, the policy is bounded by the *underlying reward models*; misjudged relevance could induce undesired behaviors. Third, our evaluation uses *exact-match QA*, which does not account for the uncertainty or multiple valid answers common in real-world scenarios. Finally, we focus on *tagged accuracy*, leaving the quality of internal reasoning spans for future evaluation.

Future work should extend refusal training to tasks like dialogue, summarization, and code generation. Integrating uncertainty estimation with reward shaping could further refine refusal calibration. Additionally, scaling the approach and testing alternative, less brittle output formats would clarify RSRR's robustness and benefits in less constrained settings.

Acknowledgments

This research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.-183, and KICH3.LTP.20.006, and the European Union under grant agreement No. 101201510 (UNITE). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905* (2024).

[2] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737* (2025).

[3] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. *npj Digital Medicine* 8, 1 (2025), 274.

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *International Conference on Machine Learning*. PMLR, 2206–2240.

[5] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* 9 (2021), 346–361.

[6] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *International Conference on Machine Learning*. PMLR, 3929–3938.

[7] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[8] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.

[9] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2567–2577.

[10] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A Manually Curated Corpus for Biomedical Question Answering. *Scientific Data* 10, 1 (2023), 170.

[11] Jaedong Lee, Hyosoung Cha, Yul Hwangbo, and Wonjoong Cheon. 2024. Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines. *Journal of Personalized Medicine* 14, 12 (2024), 1131.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[13] Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2021. Learning Cross-Lingual IR from an English Retriever. *arXiv preprint arXiv:2112.08185* (2021).

[14] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the Alignment Tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). ACL, Miami, Florida, USA, 580–606. doi:10.18653/v1/2024.emnlp-main.35

[15] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451* (2024).

[16] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. ACM, New York, NY, USA, 2421–2425. doi:10.1145/3626772.3657951

[17] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule Based Rewards for Language Model Safety. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.

[18] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332* (2022).

[19] Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2024. Reward-RAG: Enhancing RAG with Reward Driven Supervision. *arXiv preprint arXiv:2410.03780* (2024).

[20] Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. *arXiv preprint arXiv:2502.14302* (2025).

[21] Peng Qi, Haejun Lee, Oghenetegiri "TG" Sido, and Christopher D. Manning. 2020. Retrieve, Rerank, Read, Then Iterate: Answering Open-Domain Questions of Arbitrary Complexity from Text. *arXiv preprint arXiv:2010.12527* (2020).

[22] Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. 2025. Equilibrate RLHF: Towards Balancing Helpfulness-Safety Trade-off in Large Language Models. *arXiv preprint arXiv:2502.11555* (2025).

[23] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient Large Language Models: A Survey. *arXiv preprint arXiv:2312.03863* (2024).

[24] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2025. Reinforcement Learning Enhanced LLMs: A Survey. *arXiv preprint arXiv:2412.10400* (2025).

[25] Xubin Wang, Zhiqing Tang, Jianxiong Guo, Tianhui Meng, Chenhao Wang, Tian Wang, and Weijia Jia. 2025. Empowering Edge Intelligence: A Comprehensive Survey on on-Device Ai Models. *Comput. Surveys* 57, 9 (2025), 1–39.

[26] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable Visual Question Answering: Abstain Rather than Answer Incorrectly. In *European Conference on Computer Vision*. Springer, 148–166.

[27] Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection Improves Reliability: Training Llms to Refuse Unknown Questions Using RL from Knowledge Feedback. *arXiv preprint arXiv:2403.18349* (2024).

[28] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-Device Language Models: A Comprehensive Review. *arXiv preprint arXiv:2409.00088* (2024).

[29] Hanning Zhang, Juntong Song, Juno Zhu, Yuanhao Wu, Tong Zhang, and Cheng Niu. 2025. RAG-reward: Optimizing RAG with Reward Modeling and RLHF. *arXiv preprint arXiv:2501.13264* (2025).

[30] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An Open-Source Small Language Model. *arXiv preprint arXiv:2401.02385* (2024).

[31] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593* (2020).